Author's response to reviews

Title: A re-evaluation of the scratch test for locating the liver edge: a reliable physical sign

Authors:

Krishan Gupta (krishan.gupta@hnehealth.nsw.gov.au)
Akash Dhawan (akash.dhawan@hnehealth.nsw.gov.au)
Christian Abel (christian.abel@hnehealth.nsw.gov.au)
Nicholas Talley (nicholas.talley@newcastle.edu.au)
John Attia (john.attia@newcastle.edu.au)

Version: 2 Date: 15 January 2013

Author's response to reviews: see over
Response to reviewers

Reviewer 1 report:

1. I have no concern with the methods clinically and statistically used in the paper. The abbreviation “ICC” has to be explained before first appearance in the text (see “Results” in the abstract)

   **Done**

2. The term “body habitus” sounds imprecisely and should be corrected. The cut-off p-value considered as statistically significant and the tests used to compare patients (see also comments concerning results section) should be provided.

   *Body habitus has been changed to body mass index and threshold p-value has been added in the methods section.*

3. Were the patients continuously breathing during the examinations or stopped they breathing during measurements?

   *Patients were asked to hold their breath during the ultrasound exam but not during the scratch test in order to mimic usual clinical practice; this has been clarified in the methods section.*

4. The authors should clearly determinate “underestimation” and furthermore “overestimation” in relation to the reference point (below/above/nearer to the pelvis or the clavicle).

   *We have added a clarification to this effect.*

5. The authors should provide the main biometric statistics concerning the patients included in both parts of the study (e.g. age, male/female, height, weight and the resulting body mass index).

   *Table has been added and labeled table 1. The original table 1 now becomes table 2.*

6. Furthermore, authors have to describe the differences between the 15 patients examined by Rater 1 and the 16 patients examinded by Rater 2 using standard statistical tests.

   *This has been addressed by adding the table, as described above.*

7. The y-axis on table 2 is missing (“Rater 1”, isn’t it?)

   *This may be an error on the reviewer’s screen. Figure 2 is already labeled with rater 1 and rater 2.*

8. Authors postulate that a Spearmans rho of 0.37 shows only a moderate
agreement. That is generally right. I would suppose to underline that statement with the corresponding literature.

_This is a rule of thumb among statisticians and does not have a citable reference that we can find._

9. The y-axis of figure 3 (“...raters pooled”) doesn’t correspond to that what is written in the text (“...difference between a rater”). Please correct or comment on that.

_The y-axis of figure 3 shows the difference between the rater and the ultrasound; where there is more than 1 rater we use the pooled raters vs the ultrasound value. We have added this clarification in the text._

10. The calculation of ratings within the 1,2 or 3 cm of the reference value has to be described in the methods section. How did you define 1 cm, 2 cm and 3 cm. Perhaps 0.5 cm/2cm and 1.5 cm above and below the reference value? Please comment on that, as written in the paper, it sounds somehow misleading.

_As discussed in the results, “1 cm” refers to ratings that are between 0 and 1 cm off the ultrasound value; “2 cm” refers to ratings that are between 0 and 2 cm off the ultrasound value; and “3 cm” refers to ratings that are between 0 and 3 cm off the ultrasound value._

11. Figure 4 should be omitted. Instead of that figure, a second Bland Altman Graph should be provided comparing twq BMI-groups at a cut-off value (set at the median BMI).

_We respectfully disagree. Instead of comparing only 2 BMI groups and having to justify the choice of these 2 groups we prefer to show the entire range of BMIs; in this way the reader can see the “dose-response” curve of increasing discrepancy between rater and ultrasound with increasing BMI._

12. Linear regression analysis should be provided as a table with RR (alternatively : exp (beta)), 95% confidence interval and p-value.

_We have added this table of regression coefficients with figure 4._

Discussion:

1. The authors should discuss more extensively their paper concerning clinical relevance of knowing the lower liver edge in the midclavicular line.

_This is obviously the first step in judging whether there is hepatomegaly or not and we already describe this in the introduction and discussion._

2. The fact that physicians tend to underestimate objects and intraoperative findings is well known. The authors should cite some recent papers and discuss
their results in that context.

*We are referring to the fact that as the liver edge nears the right costal margin or the iliac crest it becomes more difficult to judge the exact distance; we do not see how this relates to underestimation (in the sense of minimization of findings) in other contexts.*

3. After having read the article, the reader gets the impression (especially based on the statistical tests) that the scratch test is reliable but not precise. The authors should include this fact into the title of their paper. “Re-evaluation.....: reliable but imprecise.”).

*We agree that the scratch test is reliable but would argue that it is precise enough to be used in the clinical context depending on how close one would like to be compared to ultrasound guidance. We realize that precision in this context is debatable and would prefer not to put this in the title.*

4. The authors should discuss the limitations of their study; small number of patients included in the study and the natural bias which may be given by the rater itself (degree of hearing loss etc.)

*Our study is the largest to date of this test and we already discuss extensively how it relates to other studies in this area.*

5. The authors should discuss their results in the context of further medical education for students

*Until these results are replicated we believe it may be premature to make a decision on what to teach medical students.*

**Reviewer 2 report:**

**Major Revisions:**

1. Would like to see more discussion of how this paper differs from the papers in J Clin Gastro and Lancet, which both seemed to put the nail in the coffin of the scratch test. I enjoy studies such as these, but given what has been published already, the arguments will need to be quite convincing to swing my opinion the other way.

*We already highlight the source of the differences between our study and previous studies in the discussion section; these include problems with previous studies such as:*

- Lack of standardization
- Comparing scratch test to total liver span rather than liver edge, which brings in another source of variation in locating the upper (diaphragmatic) edge of the liver
- Variability in locating the mid-clavicular line
- Not being blinded to results of other tests of hepatomegaly, e.g. palpation
- Lack of detail about how the scratch test was carried out
- Small numbers of patients

2. Comment more on diaphragm height and how this will influence the distance of the liver below the CM.

_We have already cited this as an extra source of variation in previous studies and have included all the previous literature that we can find._

3. You chose a number of tests for looking at agreement and would like to hear more justification of why these particular tests were chosen and not just a single test. Introducing so many tests introduces the bias of repeated tests; with enough tests, one will often become significant.

_The main test of agreement is that between raters and between raters and ultrasound. The other tests are secondary and explore the source and degree of this agreement. We have clarified this in the text._

Minor revisions:

1. Why was Kappa not used (takes into account agreement beyond that which would be obtained by chance alone)?

_**Kappa would not be appropriate here since kappa is for a dichotomous outcome (presence or absence of a sign) and we have continuous values (cm by rater vs cm by ultrasound).**_

I did like the way you tested for the possibility of skin transmission vs liver transmission and documented this well.

2. It would be good to document what should be considered a clinically significant distance below the RCM; I was taught that palpating or percussing the liver edge more than 3 cm below the CM meant we could call this liver 'hepatomegally' and apply a standard set of differentials to this. Yet the scratch test is within 3 cm of the US measurement (different 3 cm, I know) only just over half the time. To me this might call into question the utility of the test altogether- would need to justify this to me or explain it (in case I am misinterpreting your data).

_It is not possible to say that more than 3 cm below the costal margin should be considered hepatomegaly because it all depends on the location of the upper liver edge (overall span) and this can be very variable. This is_
the mistake that previous studies have made and incorporating this degree of error makes the scratch test look falsely poor, i.e. the scratch test could be completely accurate in estimating the lower liver edge but because the upper edge is so variable the test still looks poor. This is addressed in the discussion.